On the Use of Semantic Annotations for Supporting Provenance in Grids

Liming Chen[1], Zhuoan Jiao[2], and Simon J Cox[2]

[1]School of Computing and Mathematics
University of Ulster, Newtownabbey, Co. Antrim, BT37 0QB, U.K.
l.chen@ulster.ac.uk
[2]School of Engineering Sciences
University of Southampton, Southampton SO17 1BJ, UK
{z.jiao, s.j.cox}@soton.ac.uk

Abstract. There has seen a strong demand for provenance in grid applications, which enables users to trace how a particular result has been arrived at by identifying the resources, configurations and execution settings. In this paper we analyses the requirements of provenance support and discusses the nature and characteristics of provenance data on the Grid. We define a new conception called augmented provenance that enhances conventional provenance data with extensive metadata and semantics. A hybrid approach is proposed for the creation and management of augmented provenance in which semantic annotation is used to generate semantic provenance data and the database management system is used for execution data management. The approach has been applied to a real world application, and tools and GUIs are developed to facilitate provenance management and exploitation.

1 Introduction

The essence of Grid computing is the sharing and reuse of distributed, heterogeneous resources for coordinated problem solving in dynamic, multi-institutional virtual organizations (VO). In service-oriented grid infrastructures such as OGSA [1] and WSRF [2], grid resources are regarded as services, and problem solving amounts to the discovery and composition of the required services into a workflow, plus the enactment of the workflow. Problem solving on the Grid is dynamic, collaborative and distributed, e.g. VOs are formed or disbanded on-demand, and services may be published and withdrawn by different stakeholders. In such dynamic environments, it is vital to record the problem solving process for later use such as in interpreting results, verifying that the correct process took place or tracing where data came from.

There has seen an increasing demand for provenance in grid applications [3], which enables users to trace how a particular result has been obtained by identifying the resources, configurations and execution settings. However, current grid architectures lack approaches, mechanisms, and tools to deal with this issue. In this paper we analyse the requirements of provenance support and discuss the nature and characteristics of provenance data on the Grid. We define a new conception called

augmented provenance that enhances conventional provenance data with extensive metadata and semantics. We propose a hybrid approach for the creation and management of augmented provenance by exploiting the emerging Semantic Web technologies and the latest database technologies. The cornerstone of the approach is the use of ontologies for metadata modeling, and semantic annotations for provenance data population. Special emphasis is placed on semantics, i.e. the ontological relationships among the diversity of provenance data, which enables deep use of provenance data by reasoning.

The paper is organized as follows: Section 2 introduces the concept of augmented provenance. Section 3 describes a hybrid approach for recording and managing augmented provenance. We give an application example in Section 4, and discuss related work and our experience in Section 5. Section 6 concludes the paper and points out some future work.

2 Augmented Provenance

Provenance is defined, in the Oxford English Dictionary, as (i) the fact of coming from some particular source, origin, derivation; (ii) the history or pedigree of a work of art, manuscript, rare book, etc. This definition regards provenance as the derivation from a particular source to a specific state of an item, which particularly refers to physical objects. For example, in museum and archive management a collection is required to have archival history regarding its acquisition, ownership and custody.

In the context of Grid computing, we focus on electronic data produced by computer systems, and we define the provenance of a piece of data as the process that led to that piece of data [4]. A process in the service-oriented grid architecture refers to the execution of a workflow, which is a specification of a service composition. Therefore, the provenance of a piece of data is, in essence, the description of the process that resulted in that data item.

Grids have the characteristics of dynamic provisioning and across-institutional sharing. In such environments a workflow consists of services from multiple organizations in a dynamic VO. The success of workflow execution depends on domain knowledge for service selection and configuration, and mutual understanding of service providers and consumers on service functionalities and execution. The complexity of problem solving process requires not only the execution data of a workflow (e.g. the inputs and outputs of services, the configuration of service control parameters), but also rich metadata data about the services themselves (e.g. their usages, the runtime environment setting, etc.), in order to validate, repeat and further investigate the problem solving process at a later stage. A number of requirements for provenance data are identified and described below.

Firstly, provenance should include metadata at multiple levels of abstraction, i.e. process level, service level and data level. For example, an instantiated workflow instance is a provenance record for the data derived/generated from it, but the workflow instance itself also needs provenance information, e.g. the workflow specification it was instantiated from, the reason a particular set of input values were chosen, etc. Similar provenance requirement applies to services and data.

Secondly, provenance should include metadata from multiple categories including data, knowledge, decision, conclusion, etc. Each category of provenance has its roles and uses, and different applications have different emphases and requirements for provenance. For instance, in biology attention is paid on the transformation process of data; in engineering the focus is on the process creation; and in medical information system the emphasis is on the underlying decision-making process and results that may be more relevant to annotation. As provenance is not only used to validate, repeat and analyze previous executions but more importantly to further advance investigation and exploration based on present results, we are particularly interested in the knowledge and decision provenance, e.g. how a decision was arrived at.

Thirdly, provenance data should be interoperable, accessible and machine processable for sharing among distributed users. This requires provenance data and rich relationships among them be formally modeled and represented. Relations can be regarded as a kind of knowledge model and be used to encode domain knowledge. Appropriate organization of metadata help data retrieval and more importantly, discovery of new knowledge or pattern based on reasoning.

To meet the aforementioned requirements, we face two challenges: the first is how to capture all provenance data. While it is desirable to collect provenance data automatically, it becomes clear that not all provenance data can be captured automatically, especially regarding the rich metadata about services, workflows, knowledge and decisions. The second challenge is how to make provenance data interoperable, sharable and understandable for both humans and machines on the Grid.

Based on the above analysis and inspired by the Semantic Web technologies, we argue that ontologies and semantic annotation should be used for the acquisition, modeling, representation and reuse of provenance data. The reasons are (1) ontologies can model both provenance data and their contexts in an unambiguous way; (2) provenance data generated via semantic annotation are accessible, shareable and machine processable on the Grid; and (3) the Semantic Web technologies and infrastructure can be exploited to facilitate provenance data acquisition, representation, storage and reasoning. For example, it is straightforward to adopt Semantic Web Services for capturing the semantic metadata.

To differentiate from traditional provenance understanding, we introduce the concept of *augmented provenance*, defined as: the augment provenance of a piece of data is the process that leads to the data and its related semantic metadata.

3 A Hybrid Approach to Augmented Provenance

Augmented provenance contains execution data, e.g. the values of inputs and outputs of services; as well as semantic metadata, e.g. the descriptive information about the workflows, services and parameters. The different nature of these two types of data are reflected in the way they are captured, modeled, represented and stored. To support the heterogeneity of provenance data on the Grid a hybrid approach is proposed, which combines the emerging Semantic Web technologies with the database technologies to handle a workflow's semantic metadata and execution data respectively. The overall architecture is illustrated in Figure 1.

3.1 Managing Semantic Metadata

Managing semantic metadata for augmented provenance involves the metadata creation, semantic enrichment, representation and storage. By using the Semantic Web technologies, our idea is to formally model the semantic metadata in ontologies, thus their creation and enrichment can be accomplished in one process through semantic annotations. The generated metadata can be represented in semantic web languages such as RDF or OWL[1], and stored in semantic repositories such as 3Store [5] or Instance Store [6].

The above idea is realized in the architecture by a number of components, namely the Services, Ontologies, Semantic Metadata Repositories, Workflow Construction Environment and Query Tools. Central to the architecture is the Ontologies component containing various domain-related ontologies that specify ontological concepts, their relationships and constraints.

The Services component consists of distributed, internet-accessible services. Such services are generally described in WSDL[1] published in UDDI[2] and invoked by SOAP[1]. However these technologies do not provide formal support for service metadata and semantics. Our approach is to generate the service-level semantic metadata by semantically annotating services using ontologies, and store them in the Semantic Metadata Repositories. Composing services into a workflow is performed in the Workflow Construction Environment component. Service semantic metadata are linked to the workflow and the overall semantic metadata about the workflow are created through semantic annotations and stored in the Semantic Metadata Repositories as well.

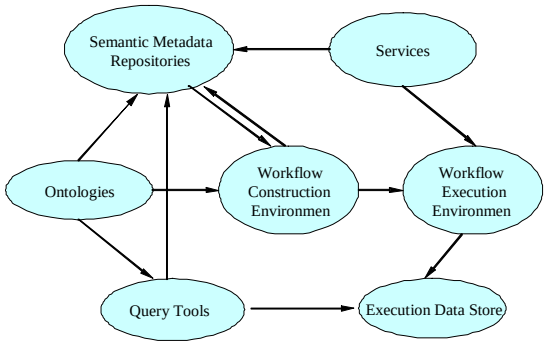

Fig. 1. The architecture for augmented provenance

The Query Tools component is for finding the required semantic metadata and execution data of the augmented provenance, as discussed later.

3.2 Managing Execution Data

Execution data include the input/output values of services, values of services control parameters, and data produced by the workflow. They have the nature of few metadata and semantics attached, but large in volume. For example, the simulation result of an aero-engine design could reach multi-gigabytes in size. Therefore, we

[1] RDF, OWL, WSDL and SOAP are W3C standards. Please refer to www.w3.org
[2] UDDI: www.uddi.org

leverage database technologies in the Execution Data Store component to facilitate the execution data storage and retrieval.

The Workflow Execution Environment component is responsible for extracting the execution data from the workflow before executing it. It analyses a workflow script to collect initial default or user-defined input values. During the runtime it interprets the workflow script and binds individual constituent services with corresponding inputs and invokes the service. Intermediate results may be returned to the environment and used as inputs to the successive services. The collected and generated data are archived in the Execution Data Store.

3.3 Querying Augmented Provenance Data

Augmented provenance consists of semantic metadata and execution data, and they are represented and managed using different mechanisms. However, semantic metadata and execution data are closely linked and can be cross-referenced. When a workflow template is built with attached semantic metadata in the workflow construction environment, it is stored in the Semantic Metadata Repositories, together with a specifically generated unique ID (UUID, Universally Unique IDentifier [7]) as a handle for later reference. An instantiated workflow template creates a workflow instance which is executed in the Workflow Execution Environment. The executable workflow instance is stored, under its own unique ID, together with associated input/output data and possibly some simple metadata (e.g. the instance creation time, name of its creator, etc) in the database. The one-to-many relationships between the workflow template ID and the workflow instance IDs are also stored in the database, so that users can reference the semantic metadata of the workflow instances through the workflow template ID.

We have implemented the Query Tools component to provide dual query mechanisms for flexible and efficient provenance data search and retrieval. Semantic queries on workflows can be framed using ontologies and are answered through semantic matching. Once a workflow template ID becomes available, its executable instances can be found easily based on the ID by launching a database query.

The separation of semantic metadata and execution data has the following advantages: Firstly, semantic metadata can be formally modeled using ontologies and represented in expressive web ontology languages. This helps capture domain knowledge and enhance interoperability. Secondly, workflow execution usually produces large volume of data that have little added value for reasoning, but storing them in the database made the data searchable and easy to share. Finally, the hybrid query mechanism provides flexibility and alternatives – users can perform semantics based query or direct database query or a combination of the two to meet application requirements.

4 GEODISE: A Case Study of Augmented Provenance

Engineering Design Search and Optimisation (EDSO) is a computationally and data intensive process whereby existing engineering modeling and analysis capabilities are

exploited to yield improved designs. An EDSO process usually comprises many different tasks. For example, the design optimization of an aero-engine or wing may involve: specify the wing geometry in a parametric form; generate a mesh for the design; decide which analysis code to use and carry out the analysis; decide the optimisation schedule; and finally execute the optimisation run coupled to the analysis code. Apparently a problem solving process in EDSO is a process of constructing and executing a workflow.

The Grid Enabled Optimisation and Design Search in Engineering (GEODISE) project [8] aims to aid engineers in the EDSO process by providing a range of Grid services comprising a suite of design optimization and search tools, computation packages, data management tools, analysis and knowledge resources. Additionally, GEODISE also intends to manage design provenance so that previous designs can be validated, repeated and further explored to lead to better designs.

We have applied the proposed hybrid approach for augmented provenance in GEODISE to help engineers answer provenance-related questions in the design process. Figure 2 shows the provenance management system.

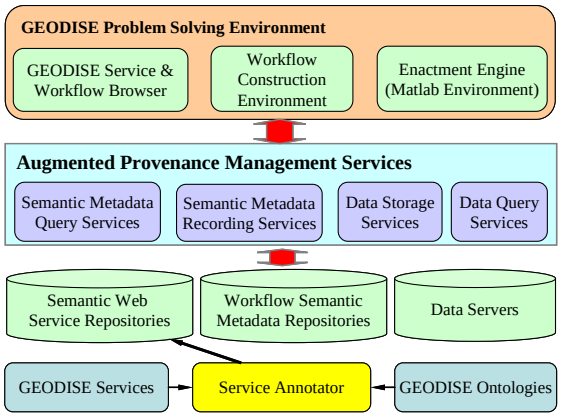

Fig. 2. Augmented provenance management system

To formally model EDSO metadata, we have developed GEODISE domain ontology and service ontology. We regard a workflow as a composite service, therefore, the service ontology can be used for modelling both service and workflow metadata. The GEODISE service ontology is based on OWL-S [9] upper service ontology which is an OWL-based Web Service ontology. It further extends OWL-S by incorporating EDSO specific metadata such as *algorithmUsed, previousService, followingService, derivedFrom, leadTo,* etc., as shown in Figure 3. The left column displays the main concepts while the right column lists concept properties.

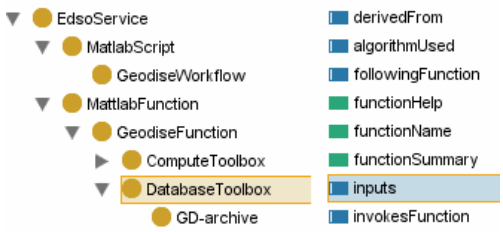

Fig. 3. An example of GEODISE service

Semantic metadata annotation API is developed for capturing augmented provenance data [10] [11]. A front-end GUI is provided to help users enrich the automatically extracted service metadata using EDSO domain and service ontologies. The annotation API is also used to capture and annotate workflow metadata during workflow construction. The generated semantic metadata

for both services and workflows are represented in OWL and stored in the Semantic Metadata Repositories, implemented using the Instance Store technology [6].

The execution data are managed by the GEODISE database toolbox [13]. The database toolbox exposes its data management capabilities to the client applications through Java API, as well as a set of Matlab functions. The Java API has been used by the workflow construction environment to archive, query, and retrieve the workflow instances for reuse and sharing. As Matlab provides the workflow enactment engine in GEODISE, the toolbox's Matlab function interfaces enable data to be archived, queried and retrieved on the fly at the workflow execution time. Data related to a workflow instance are logically grouped together using the *datagroup* mechanism supported by the database toolbox.

Querying augmented provenance in GEODISE is supported through semantic and database query tools, as shown in Figure 4. The semantic query GUI utilises the description logic based reasoning engine Racer [12] to reason over semantic metadata, and the construction of query expressions are supported by the service ontology. Here are two examples of using the dual query mechanism:

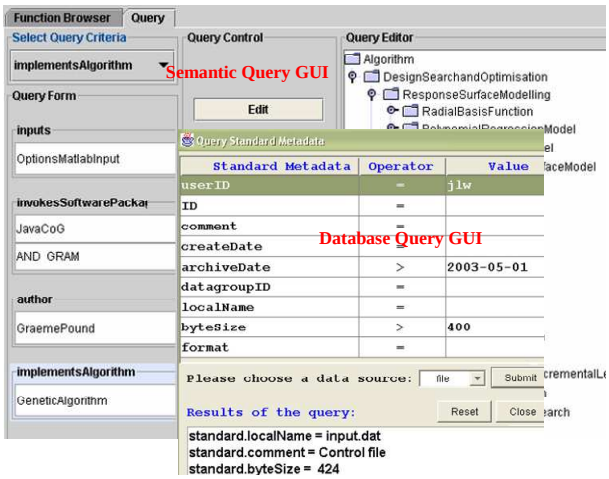

Fig. 4. Query GUIs for augmented provenance

- Find the data derivation pathway for a given design result. Actions: querying the database to find the workflow instance that is responsible for the result. Additional semantic metadata about the workflow instance can be obtained using the Semantic query GUI based on the workflow template ID. The retrieved workflow script can be enacted in the enactment engine (Matlab) for a re-run if necessary.
- Find information about the optimisation service used in the workflow that generates the given result. Actions: based on the above search, the workflow template ID is available and can be used in the Semantic query GUI to find the information about the optimisation service used in the workflow.

We have also wrapped the semantic query functionalities as web services, thus making the provenance management system easy to be integrated into service-oriented grid applications.

5 Related Work and Discussion

Provenance has traditionally been used and explored in museum, library and archival management systems where it is mainly referred to the acquisition and creation information, and the history of the ownership and custody of a resource. Research on provenance of computer-generated data has been conducted under different banners, including audit trail, lineage, dataset dependence and execution trace. Such research is mainly undertaken in domain specific applications such as geographic information system [15] and satellite image processing [16]. The common features of Chimera Virtual Data System [17], CCLRC metadata manager [18] and systems developed in [19][20] are that they try to trace the movement of data between data sources and obtain information on the *where* and *why* of a data item of interest as a result of a database operation.

Recently research on the provenance of service-based problem solving processes has attracted more attention with the prevalence of service-oriented computing paradigm. An initial attempt has been made in myGrid project [21] in which derivation provenance (log files) has been annotated and recorded for experiment validation and recreation [22]. Other systems supporting provenance include the Scientific Application Middleware [23] and the e-notebook [24]. An on-going systematic research is also conducted in EU PROVENANCE project which aims to develop a generic architecture for service-oriented provenance system [4] [25]. It also intends to propose protocols and standards to formally standardize provenance computing in service-oriented architecture.

Our work differs from the previous work in two aspects: Firstly we extend provenance data with rich metadata that is particularly useful in open, distributed and dynamic Grid-based problem solving environments. Secondly, we utilize the latest Semantic Web technologies for provenance metadata acquisition, modeling, representation, storage and reasoning, thus enhancing interoperability, machine processability and knowledge reuse. The hybrid approach of managing provenance data is innovative, flexible and practically easy to implement and to use.

The GEODISE case study serves several purposes: (1) it helps identify the generic characteristics of the provenance problems, and clarify user requirements in the context of service-based applications; (2) it helps to pin down the software requirements for a provenance system; (3) the successful design/implementation and operation of the provenance system have demonstrated and proved our conception of provenance, its design approaches and implementation rationale. Through the case study we have learnt two important lessons with regards to the use of provenance system, namely, tools should be provided for end users in their familiar working environments; and easy-to-use tools should hide as much technical details as possible that are not relevant to the end users.

6 Conclusions

The complexity of dynamic problem solving in service-oriented grid infrastructure requires rich semantic metadata in order to verify and further investigate previous

results. This gives rise to the conception of augmented provenance, which denotes both semantic metadata and execution data. We argue that the Semantic Web technologies, i.e. ontologies, semantic annotation, representation and storage, can be exploited for augmented provenance management. To this end, a hybrid approach is proposed together with an architecture that defines the core components and functionalities for realizing augmented provenance systems. We have developed a suite of generic APIs and front end GUIs in the context of GEODISE to implement the augmented provenance system. The approach is applicable for broader grid application domains.

The design and implementation of GEODISE provenance system is pioneering in many aspects. Firstly, the research provides a proof of concept for augmented provenance and provenance systems. Secondly, it provides guidelines towards the construction of a basic provenance system. Finally, it demonstrates a possible design and implementation pattern for provenance-enabled applications. In the future we shall focus on the seamless integration and interaction between provenance systems and domain-specific application systems, and in particular the design of a straightforward, easy-to-use query interface. We shall also futher investigate the security and scalability issues.

Acknowledgement

This work is supported by the UK EPSRC GEODISE e-Science pilot project (GR/R67705/01). The authors gratefully acknowledge the contributions from and discussion with EPSRC projects myGrid (GR/R67743/01) and AKT (GR/N15764/01(P)).

References

1. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: Grid Services for Distributed System Integration, Computer, 35(6), 37-46, (2002)
2. WSRF: www.globus.org/wsrf/
3. Fox, G. and Walker, D.: e-Science gap analysis, technical report, http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/GapAnalysis30June03.pdf, (2003)
4. Moreau, L., Chen, L., Groth, P., Ibbotson, J., Luck, M., Miles, M., Rana, O., Tan, V., Willmott, S. and Xu, F.: Logical architecture strawman for provenance systems, Technical report, University of Southampton. (2005)
5. Harris, S., Gibbins, N.: 3Store: Efficient Bulk RDF Storage. Proceedings of 1st International Workshop on Practical and Scalable Semantic Systems, Florida, USA, 1-15. (2003)
6. Horrocks, I., Li, L., Turi, D., Bechhofer, S.: The instance store: DL reasoning with large numbers of individuals, Proceedings of the 2004 Description Logic Workshop, BC, Canada, 31-40, (2004)
7. Mealling, M., Leach, P.J., Salz, R.: A UUID URN Namespace, IETF, October 2002. http://www.ietf.org/rfc/rfc4122.txt
8. GEODISE Project: http://www.geodise.org
9. OWL-S: www.daml.org/services/owl-s

10. Chen, L., Cox, S.J., Tao, F., Shadbolt, N.R., Goble, C., Puleston, C.: Empowering Resource Providers to Build the Semantic Grid. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), 271-278, (2004)
11. Chen, L., Shadbolt, N.R., Tao F., Goble, C.: Managing Semantic Metadata for Grid Services, International Journal of Web Service Research, (In Press), (2006)
12. Haarslev, V., Möller, R.: Racer: A Core Inference Engine for the Semantic Web, Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003), Florida, USA, 27-36, (2003)
13. Jiao, Z., Wason, J.L., Song, W., Xu, F., Eres, H., Keane, A.J., Cox, S.J.: Databases, Workflows and the Grid in a Service Oriented Environment, Euro-Par 2004, Parallel Processing, Lecture Notes in Computer Science, No.3149, 972-979.
14. Xu, F., Eres, M.H., Baker, D.J, Cox, S.J.: Tools and Support for Deploying Applications on the Grid, Proceedings of the IEEE International Conference on Services Computing (SCC 2004).
15. Lanter, D.P.: Design of a lineage-based meta-data base for GIS. Cartography and Geographic Information Systems, 18(4):255–261, (1991)
16. Frew, J., Bose, R.: Earth science workbench: A data management infrastructure for earth science products, Proceedings of the 13th International Conference on Scientific and Statistical Database Management. (2001)
17. Foster, I., Vockler, J., Wilde, M., Zhao, Y.: Chimera: A virtual data system for representing, querying, and automating data derivation, Proceedings of the 14th International Conference on Scientific and Statistical Database Management, 37-46, (2002)
18. CCLRC Data Management Group: http://www.e-science.clrc.ac.uk/web/groups/Data-Management
19. Boss, R.: A conceptual framework for composing and managing scientific data lineage, Proceedings of the 14th International Conference on Scientific and Statistical Database Management, 47-55, (2002)
20. Buneman, P., Khanna, S., Tan, W.-C.: Why and where: A characterisation of data provenance, Proceedings of the Int. Conf. on Databases Theory (ICDT), (2001)
21. myGrid project: www.mygrid.org.uk
22. Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., Greenwood, M.: Using Semantic Web Technologies for Representing e-Science Provenance, Lecturer Notes in Computer Science, No.3298, 92-106, (2004)
23. Myers, J.D., Chappell, A.R., Elder, M., Geist, A., Schwidder, J.: Reintegrating the research record, IEEE Computing in Science & Engineering, 44–50, (2003)
24. Ruth, P., Xu, D., Bhargava, B.K., Regnier, F.: E-notebook middleware for accountability and reputation based trust in distributed data sharing communities. In Proc. of 2nd Int. Conf. on Trust Management, LNCS2995, 161-175, (2004)
25. Szomszor, M., Moreau, L.: Recording and reasoning over data provenance in web and grid services, Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 03), (2003)